NPS55-83-033

NAVAL POSTGRADUATE SCHOOL

Monterey, California

A REVIEW OF SELECTED STUDIES OF COMPUTERIZED SPEECH RECOGNITION CONDUCTED AT THE NAVAL POSTGRADUATE SCHOOL

by

Douglas E. Neil

November 1983

Approved for public release; distribution unlimited

Prepared for:
l Electronics Systems Command
613
ington, DC 20360

NAVAL POSTGRADUATE SCHOOL
Monterey, California

Commodore R. H. Shumaker
Superintendent

David A. Schrady
Provost

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE *(When Data Entered)*

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER: NPS55-83-033
2. GOVT ACCESSION NO.:
3. RECIPIENT'S CATALOG NUMBER:

4. TITLE *(and Subtitle)*: A REVIEW OF SELECTED STUDIES OF COMPUTERIZED SPEECH RECOGNITION CONDUCTED AT THE NAVAL POSTGRADUATE SCHOOL

5. TYPE OF REPORT & PERIOD COVERED: Technical Report

6. PERFORMING ORG. REPORT NUMBER:

7. AUTHOR(s): Douglas E. Neil

8. CONTRACT OR GRANT NUMBER(s):

9. PERFORMING ORGANIZATION NAME AND ADDRESS: Naval Postgraduate School, Monterey, CA 93943

10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS: N0003983WRDX083

11. CONTROLLING OFFICE NAME AND ADDRESS: Naval Electronics Systems Command, Code 613, Washington, DC 20360

12. REPORT DATE: November 1983

13. NUMBER OF PAGES: 51

14. MONITORING AGENCY NAME & ADDRESS *(If different from Controlling Office)*:

15. SECURITY CLASS. *(of this report)*: Unclassified

15a. DECLASSIFICATION/DOWNGRADING SCHEDULE:

16. DISTRIBUTION STATEMENT *(of this Report)*: Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT *(of the abstract entered in Block 20, if different from Report)*:

18. SUPPLEMENTARY NOTES:

19. KEY WORDS *(Continue on reverse side if necessary and identify by block number)*:

20. ABSTRACT *(Continue on reverse side if necessary and identify by block number)*:

This report reviews selected [illegible] conducted at NPS. It includes a brief description of [illegible] and the findings. Suggestions for expansion of [illegible] and areas in which NPS has not pursued research are indicated.

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

S/N 0102-LF-014-6601

SECURITY CLASSIFICATION OF THIS PAGE *(When Data Entered)*

A REVIEW OF SELECTED STUDIES OF COMPUTERIZED SPEECH RECOGNITION CONDUCTED AT THE NAVAL POSTGRADUATE SCHOOL

ABSTRACT

This report reviews selected voice recognition experiments conducted at NPS. It includes a brief description of selected experiments and the findings. Suggestions for expansion of areas of research and areas in which NPS has not pursued research are indicated.

Rapid technological advances in Navy weapons have resulted in weapon systems with substantial increases in capability. This technological sophistication has produced the operational situation wherein the intended operator's capability to respond with the necessary speed and/or accuracy is frequently precluded. The human operator may be placed in the difficult situation where he is incapable of responding to the myriad of inputs impinging upon him. This inadequacy on the part of operators can and does result in weapon systems not reaching the full potential, or worse yet, being rendered ineffective by operator failure.

The nature of the problem is frequently one of a failure to match man's capabilities and limitations with machine system requirements. That is, machine capabilities have expanded significantly as a result of technological developments. These developments have resulted in added system complexity as well as expanded capability. However, the interface between man and machine has remained relatively constant. Man interacts with machine using the same technology that was employed with much simpler and less capable hardware systems. Therefore, while the nature of the equipment has advanced rapidly in recent years, man's capabilities and the devices provided for man to interact with machines have remained essentially unchanged. Such a situation frequently produces an environment wherein the operator cannot satisfy his function and thereby contributes to system inability to meet mission needs. Therefore, if the full potential of complex new weapon systems is to be realized,

attention must be devoted to designing man-machine interfaces which consider prospective operators in terms of system objectives.

The above is not designed to suggest that innovative techniques for man-machine communications have not been explored. For example, numerous research efforts (e.g. Connolly, 1979; Lea, 1980; Lea and Shoup, 1979; Poock, 1980; etc.) have demonstrated that speech represents a viable alternative to manual entry in man-machine interaction. These efforts have supported the hypothesis that in specific operational environments with certain task types, speech may be more effective than the traditional manual (e.g. hands, feet) control system. In fact, Lea (1980) and Martin and Welch (1980) suggest that speech, as a result of the frequency and intensity of use, is man's most "natural" and perhaps universal response modality. Further, in situations where speech can be effectively used as a human output-machine input mechanism, it may serve to free the extremities for functions incompatible with speech. Such a situation would serve to expand human operator capability and thereby enhance the possibility of the human element to function successfully in an increasingly complex and demanding military operational environment.

Based on the above, the Naval Postgraduate School began to consider the use of speech as a machine control medium in the late 1970's. The present effort represents a review of some of the major research efforts conducted at NPS and suggestions for further research in general areas of speech recognition and

speech as an input modality for machines.

Overview of NPS Voice Recognition Research

Research on voice recognition at the Naval Postgraduate School has attempted to examine the various elements which possess the potential for improving on overall system performance. The elements influencing the efficiency with which man interacts with machine include the following: (Meister 1971)

1) Equipment-physical characteristics
2) Environment-physical surroundings
3) Task-nature of job(s) performed
4) Personnel-capabilities, limitations, attitudes and training

Voice recognition research at NPS has attempted to examine the variables suggested by Meister (1971) in order to gain an appreciation for the potential effect each category may have on voice recognition performance.

In addition, it must be recognized that the arrangement or organization of the above variable categories represents a system. Therefore, in addition to considering individual parameters it is essential the combination of elements also be considered. To accomplish this aspect, the NPS research program on voice recognition included research efforts directed at performance assessment of simulated operational environments to examine combinations of variable categories.

The present effort represents a review of completed voice recognition research efforts conducted at NPS. It will include a

brief summary of each effort and a discussion of research work, as suggested by current findings. The approach taken here will begin with a discussion of student studies of individual elements followed by combination of efforts where multiple elements in a simulated operational situation were investigated.

The voice recognition system used in the majority of NPS voice recognition experiements consisted of model T600 Threshold Technology, Inc., a voice recognition system. This system is a discrete utterance recognizer. (An utterance being defined as a single word or continuous string of words not exceeding two seconds in duration.) The T600 consists of a noise cancelling microphone, analog speech preprocessor, microcomputer, CRT and keyboard unit and a magnetic tape cartridge unit. The system operates by having a subject establish reference speech patterns for a specific vocabulary during a training period. Training consists of having a subject "train" the voice recognizer by repeating each utterance to be used ten times. Training provides a basis for comparison during operation. Training can be accomplished in less than ten exposures, however, the manufacturer has suggested ten repetitions for maximum recognition. Following training, an operator speaks the utterance into the microphone followed by acceptance of the utterance by the speech processor. The speech processor extracts speech parameters from the input and converts them to digital signals for processing by microcomputer. Possible responses by the system consist of a match between the spoken utterance and the stored vocabulary; a misrecognition (i.e. systems failing to accurately recognize the spoken utterance);

non-recognition in which the system responds with a "beep".

A more detailed description of the T600 and the operational characteristics can be obtained in Armstrong (1980).

Environment-physical Surroundings

In any situation involving speech communication, noise represents a potentially disrupting influence. This potential is particularly important in situations where accurate communication is critical (McCormick and Sanders, 1982). Accuracy of reception in the presence of noise is an essential criteria for any input system being considered by the military. The operational military environment is frequently characterized by high noise levels and investigation of the potential input of noise on voice recognition must be considered of merit.

Elster (1980) conducted a study where consideration of the potential input of environmental noise on speech recognition was investigated. Elster's experiment used the T600 voice recognition system expanded to 256 two second utterances. In Elster's experiment, however, he limited the investigation to 50 utterances.

Independent variables consisted of a) noise level during training of the system, and b) noise level during testing of the system. Noise levels for both independent variables were identical and consisted of: Ambient noise (average of 38 dBA), conversational noise (average of 65 dBA), and a second conversational noise level (average of 75 dBA). In all three noise conditions deviation from the average did not exceed ± 7 dBA.

Each subject in the study trained the T600 under one of the above described noise conditions and tested in each of the three noise conditions, six subjects being randomly assigned to each training condition.

Procedure followed required each subject to train the system on the selected 50 utterances. Training was considered acceptable when an utterance was correctly recognized two out of three times. Testing of the system required each subject to voice each utterance once under each noise condition. An error was scored if the voice recognizer responded with a "beep", the wrong word, or produced output when no utterance was produced by the subject. In a second analysis of the data, Elster removed the variable of "no utterance" leaving only "beeps" and wrong responses. However, in the final analysis of variance, removal of "no utterance" did not influence the results.

Overall results of Elster's effort suggested that the presence of noise during testing influenced performance, as measured by the experiment. Specifically, Elster observed that a conversational noise background of 75 dBA produced more errors than either 38 dBA or 65 dBA. However, no difference was observed between 38 and 65 dBA. Elster did not observe a relationship between training and testing background noise levels.

This observation did not support the findings of Drennen (1980) as reported during a DOD sponsored conference on voice interactive systems. Drennen reported an interaction between

training background noise and testing background noise levels. Specifically, Drennen observed testing performance was enhanced when training occurred at noise levels similar to what was experienced during testing. The difference between findings of Elster and Drennen could reside in the fact that Drennen's noise levels were considerably higher in intensity (100 dB as measured on an SPL meter) than those employed by Elster. In fact Drennen is of the opinion interaction between testing and training will not be observed until background dB levels approach 100 dB.

The efforts of Elster were significant in that they produced evidence of potential performance degradation of voice recognition systems under consistent background noise levels. This work needs to be expanded to examine more extreme levels of background noise and different types of noise (i.e. different noise sources). For example, it may be possible that machinery noise will impact differently than conversational noise. Further, the difference between Drennen and Elster on the influence of noise during training on subsequent testing should be pursued.

Elster's study investigated the most obvious environmental influences -- noise, specifically conversational noise. As suggested above, the study should be expanded to investigate other noise sources (e.g. impact and machinery noises, etc.) as well as other noise parameters and their effect on speech recognition system performance.

However, it will be interesting and in fact necessary before implementation to investigate the potential impact of other

environmental factors. For example, vibration and acceleration/deceleration, pressure variation, etc. found in many military situations. These factors as well as any other potential environmental influences that may exist in the work station need to be considered prior to acceptance or rejection of the system.

Therefore, it must be concluded that considerable research on environmental factors and their influence on system performance remain to be done. Elster's efforts represent an excellent beginning in the area of constant noise but should be followed with additional studies in the total area of environment.

Task Variables

Another area of interest is the nature of the task to be performed. The Naval Postgraduate School research program has devoted considerable attention to the influence task difference may exercise on overall system performance. These efforts have concentrated on attempting to simulate operational type tasks.

Jay (1981) considered speech recognition as a means of improving speed and reliability in the intelligence community task of imagery interpretation. Military imagery interpretation is essentially the analysis of a display in terms of "what", "who", "when" and "where". It is designed to aid the commander in the decision making process and is a major factor in command and control. Current systems provide for man-machine communication through the use of a keyboard. Jay investigated the possibility of improving imagery interpretation by improving

the man-machine interface. It was the opinion of Jay, that improving the interaction would result in improved use of man's skills by allowing the operator to concentrate on image analysis as opposed to concentrating on inputting information into the system via a keyboard.

The research effort was designed to determine whether or not a currently available voice recognition system could be employed for reporting imagery derived from intelligence information using an interactive computer system. The question to be answered involved a determination of any significant differences in speed, accuracy, efficiency and subject attitudes regarding manual keyboard input methods versus voice entry.

Equipment employed in the study included the T600 Voice recognizer and a G 1130 Harvard Tachistoscope for simulating the optics portion of an imagery interpreter's task. The Tachistoscope is a sophisticated instrument which provides a mechanism for the presentation of visual information. It can be programmed to present or change stimuli at specific intervals or allow the viewer to change presented information at will. In Jay's experiment the Tachistoscope was used to simulate the visual task of imagery analysis and subsequent reporting. Thirty-six stimulus cards were used in the experiment. Content for the cards was judged on the basis of realism, even mix of ground, air and naval terms, use of USSR/Warsaw Pact vocabulary, and maintenance of a balance in number of characters in sets of stimuli.

The T600 used in the experiment had an expanded memory providing for 256 discrete utterances. In addition, two

recognition modes were used - buffered and unbuffered. In the unbuffered mode, the system outputs to the computer immediately following voice input. In the buffered mode, up to 128 utterance output strings could be stored sequentially in the buffer for subsequent output as a block of characters. Vocabulary used consisted of 255 utternaces. Included were the phonetic alphabet, numbers 0 - 25, administrative alphanumerics, special symbols and control characters, and air, ground and naval forces equipment vocabulary. In Jay's effort no attempt was made to limit the comparison set for an utterance. Rather the entire vocabular of 255 utterances was used for each spoken utterance.

Manual entry of information was accomplished by means of a keyboard. Subjects were given a typing test prior to participating in the actual experiment. Based on the results of the test, subjects were classified as either "fast" or "slow". Slow typists scores ranged from 17 to 32 words per minute (wpm) with an average of 25 wpm. Fast typists scores ranged from 33 to 58 wpm with an average of 43 wpm.

Included in the effort of Jay was consideration of interactive text editing. This feature was provided by facilities at ARPANET. Host computers in ARPANET were used to conduct and manage experimental as well as the interactive computer environment.

Subjects consisted of 20 volunteers. Of the twenty, eighteen were military and two civilian. Nineteen of the subjects were male and one was female. Most subjects (18) had observed a demonstration of the voice system. Twelve had

actually used the system in one capacity or another and 11 had researched voice for a report.

For purposes of the experiment, all subjects individually trained the system with the 255 word vocabulary. Training included orientation on proper methods in system training prior to entering utterances into memory. Following training of the system each utterance was repeated three times. Criterion for considering the system "trained" was correct two out of three training trials. Any utterance not meeting criterion was retrained.

Jay identified attitudes toward use of voice in a situation normally involving manual entry as an important variable worthy of consideration. A questionnaire was developed to assess attitudes of subjects regarding voice entry vs manual entry. Questions probed attitudes of subjects relative to acccuracy, speed, training, flexibility, etc. The questionnaire was administered before and after actual testing.

The results of Jay's efforts were impressive in supporting the possibility of using voice entry in the work environment and task type investigated. Jay observed that in reporting speed a highly significant difference ($P < .0005$) existed between explored entry modes. Jay observed that on the average, unbuffered voice condition was 41% and buffered-voice 58% faster than typed data entry. The author postulated that voice data entry allowed subjects to compose a report while simultaneously receiving information. This condition did not exist with manual entry.

Learning over trials was observed in all data entry modes. It was, however, interesting that no significant difference was

observed between fast typists and slow typists. The lack of a significant difference as a result of previous experience and competence as a typist may suggest that the typing task is sufficiently different from the task of interest here to render previous experience and capability of little value.

In the real world intelligence environment reporting accuracy would be as important if not more important than reporting speed. In terms of reporting accuracy, no significant difference existed between the conditions investigated. Fast typists, slow typists, and entry modes were not significantly different.

The third variable considered involved efficiency of reporting. In terms of efficiency, typing was considered the most efficient, (95%) buffered voice entry next (85%) and unbuffered voice least efficient (80%).

Jay suggested that efficiency differences may be the result of differences in exposure with the entry modes examined. That is, as a result of rather extensive exposure to keyboards prior to the experiment and limited voice entry exposure, keyboard was superior. This conclusion does appear to conflict with the earlier suggestion relative to the relationship of exposure and manual entry performance. This hypothesis deserves further study as it was not totally supported by the results of other aspects of the study. That is, observations with other variables (e.g. speed, accuracy) do not necessarily support Jay's conclusion. Additional exploration of the experience question would be of

great merit. It certainly suggests the need to attempt to equate experience levels in future studies.

In operational settings the question of voice entry system accuracy over time could certainly be an important consideration. The prospect of time-on-task or just elapsed time impacting on speaker performance would seem a reasonable assumption. Results, however, suggested that time did not degrade voice recognition system performance. The T600 performance in recognition accuracy over trials was 97% if only voice recognition errors were considered and 95.5% if rejects were involved.

Subject Attitudes

As suggested earlier, Jay correctly identified user attitudes as a potentially significant factor in overall system performance. As a result of the questionaire administered to determine subject attitudes, Jay considered subject attitudes to be generally positive relative to use of voice. Of particular interest was subject evaluation of speech entry following the experiment. Opinions expressed by participants were more positive at the conclusion of the effort than prior to commencement. There is no question that it is a significant advantage when potential users accept a proposed system. Rejection can lead to inefficiency and overall degration of system performance which can be eliminated only with extensive training and considerable experience. However, operator acceptance is not sufficient reason for acceptance of a system. For example, most

subjects express a definite preference for color displays and estimate their performance as being superior with the color display when compared with black and white. Research evidence does not always support the opinion of the subject in that in some tasks while perference may be for color, performance favors black and white. The point being that preference alone may not necessarily mean performance will be enhanced. This should not be construed, however, to minimize the significance of operator acceptance.

In summary, it must be admitted that Jay's study did not provide evidence of a clear superiority for manual or voice entry. The results were mixed indicating advantages in specific situations for each entry mode. Such a finding is not unique and suggests the need for additional research to identify those tasks where voice entry can provide for performance enhancement and/or identify tasks which seemingly are not well suited to voice entry.

McSorley (1981) also examined the potential of voice recognition in an applied setting. The author's objective was to examine the possibility of operating a Warfare Environmental Simulator (WES) by voice entry rather than the traditional manual method. The procedure utilized involved entry commands via voice or manual entry and subsequent evaluation of the two entry methods.

The WES wargame is computer assisted and consists of a two-sided interactive process in which two sides (blue and orange) can define, structure and control own forces. It is a naval

war game exclusively and involves 80 commands for control of platform and sensors involved in the game. Commands used by players are highly formatted in terms of syntax and input parameters. Input errors can consist of incorrect syntax or an impossible action. Syntax errors result in immediate notification which can be corrected immediately. Impossible action errors do not result in immediate warning. Notification of impossible action is not displayed until execution is attempted.

WES requires that a specific scenario be selected. Scenarios employed by McSorley consisted of the CUBA scenario because of its simplicity and adequate forces to meet objectives of the effort. In the scenario U.S. vessels consist of the aircraft carrier Enterprise, guided missile destroyer Berkley and the nuclear submarine Sturgeon. Opposition forces consisted of three Soviet warships and a merchant vessel in a situation similar to the 1962 Cuban missile crisis.

Equipment consisted of the Threshold T600 voice recognition unit, an ADM 31 Data Display Terminal and a Miniterm Model 1203 system. The T600 was used for voice entry, the ADM 31 for manual entry and the Miniterm was used to provide a hard copy printout of input commands for scoring performance.

Subjects

Twelve volunteer subjects participated in the study. Eleven of the subjects were male military officers and one was a female civilian member of the NPS faculty. Subjects had varying levels of experience with WES, with the faculty member being quite experienced with the wargame. Familiarity with the voice

recognition system also varied from experienced to inexperienced with six being assigned to each category.

Training involved use of the WES vocabulary which consisted of 162 utterances. As in the previous efforts, once the voice recognition system was trained, each vocabulary utterance was repeated three times. Utterances correctly identified two of the three times were considered "trained". Utterances failing to meet the criterion were retrained.

Typing ability was assessed using a 5 minute typing exercise. The exercise consisted of two standard paragraphs totalling 21 lines. Typing ability ranged from 20 to 40 wpm.

The actual experiment consisted of 20 basic WES commands. These commands totalled 162 utterances and involved 67 of the 162 utterances deemed necessary to conduct an actual WES war game. The 20 commands were segmented into five groups, each consisting of four commands. Subjects input the 20 commands and five groups of four commands in each of three input modes. These three input modes consisted of buffered voice, unbuffered voice and manual (typing). Input methods were randomly assigned in terms of order of presentation.

Performance measures were as follows:

(1) Time required for input

(2) Input error.

Input errors involved recognition errors and operator errors. An error involving a misrecognition by the T600, i.e. utterance was not correctly identified, was considered a recognition error. This form of input error was obviously not applicable to manual

entry. Operator error was essentially any other error that could not be classified as a recognition type error.

Results of McSorley's effort suggested that manual entry resulted in fewer errors and faster input than either buffered or unbuffered voice entry. Under unbuffered voice condition, of the 67 utterances required to form the command, 46 had been misrecognized at least once. Twenty-one of the utterances were never misrecognized. In terms of total errors, manual entry resulted in 169 errors, buffered voice 542 and unbuffered voice 701. Therefore, manual entry resulted in 68.8 percent fewer errors than buffered and 75.9 percent fewer errors than unbuffered voice.

Speed of entry also favored manual input. Total time for input using manual entry was 254.35 minutes, 286.17 for buffered and 585.7 for unbuffered. Typing was therefore 11.1 and 56.6 percent faster than buffered and unbuffered voice entry.

Experience was a definite factor in time required for entry. However, while unbuffered voice appeared to be the most dramatically affected, relative position did not change. However, experience did not impact on recognition errors. Operator errors (e.g. spelling and typing errors for manual entry and forgetting procedures in voice entry) favored buffered voice entry with no difference between manual entry and unbuffered voice.

Results of McSorley's effort seemingly favor a manual entry, particularly over unbuffered voice in the experimental task. There are, however, several considerations which require

clarification. First, error measurement was not the same for voice and manual entry. Voice entry total errors included recognition errors and operator errors whereas manual entry assessment was restricted to typing errors/operator errors. The impact of this difference in the applied sense is obviously unknown. It may be that in operational setting the fact that error measurement was not the same is really ot little importance. However it may also be that the difference does not allow the accurate assessment of the two techniques of data entry.

It may be that voice entry is simply not appropriate for tasks of the type examined by McSorley. The results support the requirement to examine the nature of the task prior to concluding that application is or is not appropriate.

Ruess (1982) studied the applicability of discrete utterance voice recognition in a simulated loading and retargeting situation for Air Launch Cruise Missiles (ACLM). As in previous efforts, Ruess compared voice entry and traditional manual (keyboard) entry of information in a retargeting situation. A second dimension, although not directly involved with entry method, was an examination of display techniques and their effect on performance.

Ruess measured speed of input, accuracy of input and time vs accuracy of entry methods. For keyboard entry, a working model of the integrated keyboard (IKB) used in the B52G was constructed. The integrated keyboard was interfaced with an Apple II microcomputer via a Lear-Siegler ADM 3A terminal. Information input via the keyboard was stored in memory for later

printout on a MICROLINE Micron 80 printer.

Voice entry was accomplished using the T600 voice recognition system. In the Ruess study the only major difference from previously observed investigation was the use of the Apple II computer.

Subjects consisted of 20 volunteers. Subject population was comprised of 16 male and four female participants. Seventeen of the 20 were military and three were civilians. Five of the subjects had previous voice entry experience and no subjects had IKB experience.

In the case of both entry methods, subjects were allowed a familiarization/training period. Voice entry training was similar to previously discussed efforts.

Experimental design involved each subject entering 20 ALCM target sites. Entry mode selected was determined using ABBA ordering technique. A second aspect of the study involved a "Target Information" verification. Purpose of this portion of the experiment was to compare display techniques. Subjects were required to make changes in certain target sets where deliberate errors had been introduced by the investigator. Certain sets required no input while others required a total of nine modifications. Each of the three groups of six sets required a total of 26 randomly distributed changes.

In "Target Loading Task" results were obtained and analyzed for time, accuracy and time vs accuracy. In this task keyboard entry was faster for 15 of the nineteen subjects. Of the four remaining subjects, one demonstrated no difference between entry

modes and three subjects had faster entry times using voice. Statistical analysis supported that IKB was faster than voice entry ($P < .05$). These results were similar to the findings of McSorley and in conflict with the findings of Jay. Overall, IKB was more accurate than voice with a $P < .01$.

In terms of output, there was no significant difference between voice and manual entry.

Overall, Ruess' experiment suggested manual entry was superior to voice in terms of speed and input accuracy. Output accuracy, the validation of data prior to actual entry, indicated no difference between entry methods.

Wolfe and Taggert (1981) compared voice and manual entry in an operational data entry task. Wolfe and Taggart wrote a computer program which would simulate data entry capabilities of the P-3C operational software.

The authors attempted to simulate an operational data input function analyzing an operational vocabulary. The input function selected for test involved the TACCO preflight data entry in the P-3C ASW patrol aircraft. The actual task involved entering preflight data in Stores Management and Navigation Preflight tableaux. In order to accomplish the task, three talbeaux from the operational software were used in the simulation. These involved: INDEX, STORES MANAGEMENT, and NAV PREFLIGHT. The INDEX tableaux represented a comprehensive representation of tableaux available to operators in the operational system. The INDEX tableaux allows operators to select the desired tableaux, in the

case of Wolfe and Taggart's experiment either STORES MANAGEMENT or NAV PREFLIGHT, by inserting the appropriate command. Once displayed, operators can interact with tableaux by means of a data entry system.

Wolfe and Taggart were interested in examining whether or not voice entry held any advantages over the traditional manual entry (keyboard) system. The authors selected speed of input and accuracy as their performance measures.

Equipment included the T600 voice system, a Datamedia Elite 2500 CRT and a keyset. The effort differed from some previous experiments in that a PDP 1150 computer was used in the experiment. Display and entry (keyset) were displaced to require subjects to divert their attention away from entry systems as is true in the operational environment.

Thirteen volunteers served as subjects. Subjects included twelve male and one female officer. All subjects had prior experience with keyboard entry with a wide range of exposure levels. Four of the subjects were TACCO's and had experience with the data entry task. One of the thirteen had previous experience with voice recognition equipment.

In preparation for participation, subjects were administered a typing test and provided with information regarding use of voice recognition equipment. Subjects were then allowed to familiarize themselves with voice reognition systems and instructed in the "training" of the system. Subjects trained the system on the 61 vocabulary utterances following familiarization.

A departure from earlier efforts was that Wolfe and Taggart set the criterion for acceptance of training at three out of four correctly recognized utterances, rather than two out of three.

Stage two of the study consisted of subjects actually filling in the data required for the STORES MANAGEMENT and NAV PREFLIGHT tableaux. Subjects had to insert data twice, once manually and once using voice. Entry method sequence was randomly assigned to subjects. Stage two followed stage one by one to three days.

Stage three followed stage two by one to three days. This consisted of revising the entry sequence. That is, subjects who started with voice in stage two, started with manual during stage three. Those who started with manual in stage two, started with voice in stage three.

As suggested earlier, Wolfe and Taggart selected speed of input and accuracy as their performance measures. Performance evaluation was based on operational error rate and time required to enter data. Operational errors were defined as entry errors which went undetected by subjects and therefore remained following completion of the data entry task. Input errors were recorded but only for consideration in analyzing overall input time.

Tableaux selected for study (i.e. STORES MANAGEMENT and NAV PREFLIGHT) were used as a result of different input requirements. STORES MANAGEMENT was selected because one utterance could provide more than one bit of data output. This was considered to be the most advantageous condition. NAV

PREFLIGHT was considered the less advantageous situation as a result of one utterance providing one bit of data.

As all subjects participated in two trials, analysis of entry time examined the effect of trials and entry method. Results indicated that in STORES MANAGEMENT, voice was faster than keyset entry in both trials. It was interesting that keyset manifested a 9.1 percent improvement in time between trial one and two, while voice entry experienced a 12.6 percent improvement in entry time between trials one and two ($P<.01$). However, statistical analysis revealed no significant interaction between entry method and trial.

In Navigation Preflight data entry a similar situation was demonstrated. Keyset data entry was 11.6 percent faster on trial two than trial one, and voice was 5.4 percent faster ($P<.15$). While not reaching the desired level of statistical significance, the findings are suggestive. Comparison of entry time for the two tableaux revealed that in STORES MANAGEMENT voice was 9.7 percent faster than manual ($P<.1$). Again the results were not significant at the desired level, however, results do suggest that for STORES MANAGEMENT voice input was faster. In NAV PREFLIGHT keyset was 14.7 percent faster than voice entry with the findings statistically significant ($P<.01$).

The findings support Wolfe and Taggart's belief that task characteristics may be a contributing factor in entry speed performance. That is, character by charcter input vs. multiple character may dictate which method of entry is superior. Between

subject variability in entry speed performance was also manifested in certain aspects of the experiment. Analysis indicated a difference between subjects on the STORES MANAGEMENT task but not NAV PREFLIGHT. Once again the data suggests that characteristics of the task may influence overall performance.

Experience level was also considered as a possible factor in entry performance. The typing test administered prior to the actual conduct of the experiment was used to segregate subjects into fast typists (30 wpm or greater) and slow typists (less than 30 wpm). No difference in entry speed was observed between the two groups.

Data had also been recorded on warfare Speciality. It will be recalled that four of the subjects had had experience with manual entry or the type of data used in the study. In the first no difference was observed between the "TACCO" group and the "non TACCO" group. However, on the second trial the "TACCO" group was 23 percent faster with a statistical significance of $P<.01$.

In terms of operational errors, (i.e., errors which remained at completion of a trial) no significant difference was observed between the entry methods. Unlike entry speed, trials did interact with performance.

Entry errors were not a prime concern of the effort. However, entry errors were considered in the analysis and differences between entry modes were observed. The error rate for voice entry was 2.4 percent and manual entry 1.2 percent. The difference was statistically significant ($P<.025$). However,

it should be noted that with the observed percentages a small shift in the absolute error rate would result in a much greater shift in the relative rate. Therefore, the findings could be misleading if strictly interpreted.

Wolfe and Taggart suggested some potential reasons for voice entry errors in their presentations. They observed, for example, that voice recognition system experienced considerable difficulty with the operational vocabulary. That is, in the vocabulary there were several utterances which were very similar (e.g., "thirteen long", "fifteen long", sixteen long", etc.). In fact eight utterances or 13 percent of the total resulted in 71.9 percent of the entry errors and five percent of the vocabulary (3 words) accounted for 41.2 percent of the errors. Elimination of these troublesome utterances would probably have improved overall performance of voice entry considerably. Certainly vocabulary selection is a variable demanding attention in a comparison of entry modes.

An interesting observation in the Wolfe and Taggart effort was the sampling of subject opinion relative to entry modes. The authors had subjects respond to a questionnaire regarding mode of entry preference. Twelve of the subjects suggested voice was the preferred entry mode in the STORES MANAGEMENT task. Reason for their preference indicating freeing the eyes for verification of input and a decrease of fatigue which they related to a decrease in the probability of producing errors. In terms of the NAV PREFLIGHT task the responses were generally neutral in terms of entry mode preference.

The overall conclusion of the effort seems to favor voice entry for the STORES MANAGEMENT task and manual entry for NAV PREFLIGHT task. These findings are significant in that they suggest a possible relationship between nature of the task and entry mode. This very important question needs further development. The tasks examined have evolved with manual entry considered as the entry method. While in some cases slightly inferior to manual entry, voice has certainly compared favorably in all cases. The question of performance effectiveness if the task had been designed with voice entry in mind as the entry mode needs to be researched before a firm conclusion of superiority can be reached.

The results of studies on operational type tasks suggest a number of areas in which further research is required.

Studies should be developed which concentrate on:

(1) nature of tasks

(2) experience levels

(3) training (e.g., criteria for suggesting system is "trained")

(4) attitudes

The data does support the possibility of voice entry in the operational control. Jay's work and Wolfe and Taggart's effort in particular suggest the value of voice entry in certain environments/tasks.

Personnel

Batchellor (1981) was concerned with the potential influence of certain personal characteristics on voice recognition system

performances. She considered sex (male vs female), officers vs enlisted, and extent of training (three, five or ten training trials).

Batchellor's study used essentially the same equipment as previously discussed efforts. Subjects were introduced to the equipment and the nature of the experiment explained. Following familiarization actual "training" of the system commenced. One objective of the effort involved consideration of the relationship between repetition of each utterance and performance. That is, the manufacturer recommends 10 training passes. However, when using 10 trials, training can be extremely time comsuming. If essentially the same results can be obtained with less training, a considerable time saving could be realized. Batchellor investigated performance with three, five and ten training trials. Order of training passes was randomized so that each (i.e., three, five and ten) was used first and with an equal number of trials. Therefore one-third of the subjects started with three training passes, one-third started with five and one-third started with ten. Batchellor used the 2 out of 3 correct recognition as her criterion for "trained".

Subjects for the study consisted of ten female officers, ten female enlisted personnel, ten male officers, and ten male enlisted personnel. Enlisted subjects were stationed at the Naval Postgraduate School.

All but two of the officer subjects were students at NPS. The remaining two consisted of an officer stationed at Fort Ord and an officer stationed at Joint Chiefs of Staff.

Vocabulary used by Batchellor consisted of 50 utterances. Utterances varied in length from one to five syllables. Criteria for selection was based on matching the number of utterances in each syllable category (i.e., having an equal number of two syllable utterances as three syllable, etc.)

Results of Batchellor's effort indicated that sex was not a major factor in performance. Machine recognition performance was slightly better for men (error rate of 1.8%) than for women (error rate of 2.1%). This difference, however, was not statistically significant. These results indicate that voice characteristics as reflected in male-female differences, do not represent a major problem in system performance.

In terms of the relationship between system performance and rank, enlisted personnel had a slightly lower mean error percentage (1.85%) than officer subjects (2.05%). This difference was not statistically significant and one can conclude that rank in and of itself did not represent a major influence on performance.

The relationship between training trials produced some very interesting results. Batchellor observed no difference between five training trials and ten training trials (1% error for both rank and sex). However, even though a significantly greater number of errors were observed with three training trials, the percentage error rate was still only three percent.

Interestingly there did appear to be a relationship between error rates and rank. Initial indications suggested enlisted performance was superior to officer with the reduced (3 training

passes) training trials. That is, there did appear to be a significant rank by number of training passes interaction. The reason for this interaction is unclear and it is possible the results were spurious. These findings do suggest the need to pursue the question of rank, and all the parameters that rank implies in relation to performance of the system. It is possible that certain characteristics of the rank structure may influence performance.

The interesting finding here, however, was the fact that within the conditions of the present experiment, little difference was observed between five and ten training trials. These findings could be of major importance and certainly merit further study. For example, does the relationship hold under an expanded vocabulary? Or can performance be maintained with fewer training trials providing the criteria for acceptance is made more rigid?

Neil and Andreason (1981) examined the bilingual capability of the T600 speech recognition system. That is, in many military situations (e.g. NATO Command and Control Center) it is possible that an operator may be required to interact with a speech recognition system in an "official" language that is different than his/her "natural" language. Even with the user that is quite proficient and fluent with the "official" language, the potential for reversion to "natural" language may be considerable under certain circumstances.

The objective of Neil and Andreason's effort was to examine the ability of the T600 to recognize utterances in either

language when training had occurred in both languages. Essentially, the effort was designed to investigate the ability of the T600 to function in a bilingual mode.

Equipment used included a T600 voice voice recognition system with additional memory modules which expanded its capability to 256 .1 to 2 second discrete utterances. In the actual experiment only 105 discrete utterances were used.

Subjects consisted of 16 volunteers; 12 males and four females. Male subjects were West German officer students at the Naval Postgraduate School. Female subjects were wives of German officer students at NPS. All subjects were bilingual (German/English) with German being the natural language in all cases. All subjects were volunteers and received no compensation for participation.

A 105 utterance list was proposed for use in the research effort. Utterances were selected on the basis of their possible application in a Command/Control type environment. No attempt was made to control for syllable count in either language, nor was any utterance accepted or rejected on the basis of its potential for accuracy in recognition.

The procedure required that each subject "train" each utterance three times. Subjects repeated each utterance 10 times in English followed by testing in English; trained each utterance 10 times in German, followed by testing in German; and repeated each utterance 5 times in English and 5 times in German followed by recognition testing in English and German. Actual order of training and testing was randomized to control

for potential interactions between training sequence and recognition performance.

Translation of English to German was performed by one of the experimenters. This was done to reduce variability in the German list. It was observed that without such control considerable variability in translation of English to German was possible.

Performance measures were considered in terms of recognition accuracy under training/testing conditions described earlier. Performance measures included misrecognition (i.e., incorrect recogniton) and non-recognition (i.e., inability of system to match test utterances with any trained utterance). Misrecognition and nonrecognition were both considered as errors and were given equal weight in analysis.

Design of the experiment was of a repeated measure type in which each subject served as his own control and was therefore tested under all conditions. The design selected allowed for determination of training effect variable and a reduction in variability associated with individual differences.

In addition, as a result of the nature of data obtained, the authors analyzed raw data and arcsin transformed data. Arcsin transformation put data into a form that would more nearly satisfy the assumption underlying analysis of variance.

Analysis of both raw and arcsin transformed data supported a highly significant training language effect. When training/testing occurred with a single language (i.e., English/English or German/German) no difference was observed. However, when training involved both languages performance was significantly

degraded. Further analysis revealed that neither language contributed a disproportionate amount to performance degradation.

In summary, the report indicated that the T600 could function equally well in either of the two languages studied (English or German) alone. However, when required to perform in a bilingual mode the variation in each utterance produced such a complex array that the T600 could not develop a satisfactory reference matrix and performance was severely degraded.

Therefore the study by Neil and Andreason suggests that any situation wherein a bilingual situation could be anticipated would almost certainly result in a reduction in recognition performance.

In any operational configuration an important consideration in voice recognition performance is time and vocabulary size. That is, if operators were required to "retrain" the system frequently when repeated use was required the time required and inconvenience created could seriously degrade the overall effectiveness and useability of the system. Obviously such a situation would be compounded with increased vocabulary.

Poock (1981) identified this potential problem area and indicated an experiment to investigate the potential for performance degradation as a function of time and vocabulary size. Subjects initially consisted of six military and two civilian. Two of the subjects were female. One male subject was forced to withdraw at the 8th week leaving a total of 7 subjects. Length of the effort was 21 weeks.

The system utilized was the Threshold Technology, Inc. Model T600 voice recognition system. Subjects observed the recommended training sequence (i.e., each subject repeated each utterance 10 times). Vocabulary consisted of 240 utterances. Following completion of training, each utterance was repeated three times. Criterion for successful training was correct recognition two out of three passes. In the event the utterance was not correctly recognized two out of three times, the utterance was retrained. Once criterion was reached, training patterns were not changed during the remaining 20 weeks of the effort.

In addition, two subjects (one male and one female) trained the T600 in a "joint mode". In "joint mode" the two subjects each trained each utterance 5 times. Same criteria for recognition sequence was adhered to and remained unchanged for the following 20 weeks of the experiment.

It should be mentioned six of the eight subjects had minimal previous exposure to the T600 voice recognition system (roughly one month). The two subjects participating in the "joint mode" function were "experienced" in that each had at least one year of experience with the system.

For the experiment, the 240 utterance list was divided into 20 utterance segments. Each segment consisted of two 1 syllable utterances, six 2 syllable utternaces, four 3 syllable utterances and four 5+ syllable utterances. The utterance list was selected from times and frequency of use experiments in a Command Center.

Actual procedure required subjects to participate each week for 20 weeks. During weekly testing each subject repeated each utterance twice. The procedure involved expanding the window by 20 utterance segments. That is, each subject was first tested only on utterances 0 - 19. Once utterance 19 was repeated the window was expanded to include uterances 0 - 39, followed by 0 - 59, etc. This was done to see if vocabulary size significantly influenced performance. The procedure allowed for examination of performance beginning with a small vocabulary (20 utterances) and expanding by 20 utterance increments up to and including the full 240 utterance list.

The two subjects selected for the "joint mode" performed as above as well as providing an additional 480 repetitions for examination of joint reference pattern performance.

At the completion of 20 weeks all subjects retrained each utterance which had been misrecognized during the 20 week testing schedule. Following retraining subjects completed testing for the 21st week.

Analysis of the results of Poock's longitudinal effort indicated time and vocabulary size were not significant factors in system performance. As expected there were between individual differences. However, the results indicated no significant within individual differences over the period of testing. In fact, over the 21 week testing period there was less than a 1.7% variation in recognition performance. The suggestion here is that reference voice patterns, over the 21 week period at least, remained very stable. Further, it will be recalled that prior to

the 21st week all utterances misrecognized during the previous 20 weeks were retrained. The normal expectation would be a significant improvement in recognition during the 21st week. However, while some slight improvement was indicated, improvement was not statically significant. Observed improvement could have just as easily resulted from "end spurt" as from retraining.

Vocabulary size did not significantly effect recognition performance either. Voice recognition remained relatively uniform as vocabulary size increased with statistical analysis indicating no increases in error rate. There was an indication that error rate was related to the number of syllables in an utterance. Increasing the number of syllables in an utterance resulted in decreased recognition performance. The suggestion here is that vocabulary size may not be a factor in performance, but that structure (syllable count) may.

One very interesting aspect of Poock's effort was the joint reference pattern investigation. Performance under joint conditions was very impressive. In fact, performance degradation was .7% when compared to their own patterns. The male subject's performance was superior to any other subject using their own individual reference patterns.

The longitudinal study conducted by Poock demonstrated that performance was not seriously degraded over time. The observed stability suggests that re-training of voice patterns may not be necessary with prolonged use. Further, the effort certainly suggests the possiblity of joint reference patterns at least for critical or "stop action" inputs.

One potentially disruptive influence in voice recognition is the concept of stress. Armstrong (1980) in a comprehenisve examination of the effects of workload on voice recognition studied the problem of task-induced stress on overall system performance.

As in previously described efforts, Armstrong employed a T600 voice recognition system. Vocabulary consisted of 50 distinct utterances. Thirty of the utterances were selected from the Modified Rhyme Test which is commonly used in the determination of speech intelligibility of communication systems. Sixteen of the 30 words actually were eight pairs of rhyming words. In each such pair the ony difference between words was the initial consonant. For example, the words "beat" and "peat" would constitute such a pair. The remaining 14 words consisted of seven pairs of non-rhyming but similar words (e.g., "sap" and "sat"). Twenty of the 50 utterances were selected by Armstrong from single words commonly used in Command and Control environments. These utterances were distinct and more easily distinguishible from the 30 utterances selected from the rhyming test.

All words used were either one or two syllables. Selection was actually based on an attempt to "confuse" the T600. this intentional confusion was attempted to demonstrate a decrement resulting from the loading task. A similar objective could have been satisfied by a considerable expansion of the vocabulary. Such an expansion would have required a considerable increase in testing time and Armstrong felt the same objective could be

accomplished by means of increasing the potential for confusion. Therefore, the vocabulary was purposely selected to increase the likelihood of recognition errors.

Subject loading was accomplished through the use of a pursuit tracker. The task involved tracking a .75 inch square light target travelling in a clockwise direction at a constant 40 rpm rate. The tracking task was made up of a circular tracking task and a square-like task. Performance was based on time on target.

Procedure consisted of a brief orientation followed by familiarization with equipment. Subjects then "trained" the 50 word vocabulary. The two out of three criterion was applied for successful training.

Experimental conditions consisted of three levels of motor loading (tracking) and the voice recognition task. In one condition, no tracking task (NTT) there was no tracking requirement. This condition assumed no motor loading. Subjects were also required to perform the circular tracking task (CTT) and the square like tracking task (STT). During the combined tracking and voice recognition pattern, it was emphasized that voice was the primary task. Presentation of tasks was presented in different orders for NTT, CTT and STT, thereby controlling for any learning or ordering effect.

In what is assumed to be an attempt to examine the effects of time on task, Armstrong had subjects repeat two different consecutive random orderings of vocabulary words. The first time through the vocabulary was considered the first half

of the trial and the second pass was referred to as the second half of the trial. Subjects were not informed as to when they had completed half of a trial.

Armstrong was also interested in the possibility that subjective fatigue may influence performance. As such, each subject was administered the "Feeling Tone" Checklist upon completion of each condition (Pearson and Byars, 1956).

Analysis consisted of an examination of (a) recognition errors, (b) subject verbal errors, (c) influence of subjective fatigue and (d) tracking performance. Recognition errors were defined as a failure of the T600 to correctly recognize a vocabulary word. This included incorrect recognition and rejection of the word as non recognition. Verbal errors were defined as the failure of a subject to correctly repeat a presented word. As suggested earlier, tracking performance was evaluated in terms of time on target. That is, the amount of time subjects were able to maintain contact with the rotating bug with their wand. Subjective fatigue was evaluated by the method suggested by Pearson and Byars (1956).

Results of Armstrong's effort suggest that loading the operator did influence recognition performance. Specifically, when all word types and both trial halves were considered the NTT resulted in an error rate of 10.51%; CTT resulted in an error rate of 14.43%, and STT resulted in an error rate of 14.73%. By trial half, including all word types and loading condition the first was slightly better 12.71% to 13.73% for the second half. Vocabulary word, overall loading conditions and both trial

halves, revealed that rhyming words had an error rate of 25.67%, non rhyming at 12.91% error rate and operational words a 3.48% error rate. Overall error rate was 13.22%.

Analysis revealed that motor loading did affect recognition performance. The difference between NTT and the loading condition of CTT and STT was significant at $P<.10$. Error rate also differed by vocabulary word type (rhyming, non-rhyming but similar and operational.) A non-parametric analysis technique indicated that pairwise comparisons of recognition error rate were significant at $P<.01$. The conclusion here being that recognition error rates for each of the vocabulary word types were different from each other word type.

Loading also influenced operator verbal performance. Not surprisingly, with increased task loading a subject's ability to repeat the stimulus word correctly was degraded. Subjective fatigue ws not found to be a significant factor in overall performance.

In summary, motor loading did negatively affect recognition and verbal performance in Armstrong's effort. The nature of the secondary tracking task was such that successful performance was extremely demanding. It was actually surprising that recognition performance and verbal errors did not suffer greater degradation. Pursuit tracking requires continuous effort on the part of a subject and is an excellent source of task-induced stress. However, it is difficult to imagine a real world situation that would require the same level of sustained attention and performance. The question of realistic motor loading and its affect on

performance is of considerable merit and should be pursued. Armstrong has shown that even though recognition performance suffered as a result of motor loading (i.e. error rates were roughly 10 times normal) the T600 performed extremely well given the nature of the task. In fact, when one considers the nature of the tracking task it should be obvious that simultaneous manual performance would be difficult if possible at all. Therefore, all things considered, the ability of subjects and equipment to function at the rather high levels observed suggests the significance of Armstrong's effort.

In a followup effort, Armstrong and Poock (1981) examined the affects of mental loading on recognition performance. The interest was directed at examining the potential relationships between increased mental load (i.e., over that experienced by subjects during training of the T600) and performance of the voice recognition system. As in Armstrong's (1980) effort, the assumption was that load may result in altered voice characteristics which would degrade overall system ability. The Poock and Armstrong (1981) effort was obviously designed to augment the work of Armstrong (1980).

The voice recognition portion of Poock and Armstrong's study was essentially the same as earlier effort of Armstrong. Vocabulary, training, equipment, etc. were basically the same for the two studies.

The loading portion of the effort was accomplished through the use of a General Dynamics Response Analysis Tester (RATER). The device is an effective instrument for investigating response

speed/accuracy as well as short term memory. In the Poock and Armstrong study the RATER was used to generate and display random sequences of four individual symbols (i.e., circle, cross, diamond and triangle). Symbols were presented at a constant rate of one symbol every 1.5 seconds. Response buttons appropriately labeled with the four symbols were provided to subjects.

An interesting feature of the RATER and one potentially valuable for the Poock and Armstrong study is the ability to program "delay" modes into the system. Delay modes enable the investigator to "delay" the proper response to the current stimulus. In other words, in delay mode zero the proper response is the currently presented stimulus. In delay mode one the proper response is the response button labeled with the previously displayed stimulus and delay mode two would be the symbol which had appeared two trials back, etc.

As such, in delay modes a subject is forced to recall stimuli presented one, two, three, etc. trials previously rather than the currently displayed information. Such a system is capable of placing a considerable mental load on subjects.

In the Poock and Armstrong effort delay condition of zero, one and two trials back as well as no mental loading were employed.

Subjects consisted of 24 volunteers. Twenty-two were male U.S. military officer-students at the Naval Postgraduate School. A female civilian and one Canadian military officer completed the subject population. Sixteen of the subjects were designated as being experienced with voice recognition equipment (2-10

hours) and eight subjects had no experience with voice recognition. Two of the 14 had had a brief (1/2 hour) experience with the RATER.

To reiterate, the hypothesis was that increased mental loading would result in changes in voice characteristics sufficient to degrade the recognition ability of the T600.

System performance was considered in terms of loading, trial half, T600 experience and vocabulary word type. Under loading Poock and Armstrong observed that with no RATER loading error rate was 10.77%; with zero delay 13.18%; with delay one 13.14% and delay two 13.60%. The suggestion here is that the only difference was between no external loading and the three loading conditions. This observation was confirmed by statistical analysis which suggested that the only significant difference existed between no loading (NRT) and the remaining three loading conditions zero delay (ND0) delay one (RD1) and delay 2 (RD2).

Recognition error rate was also observed to be higher during the first half of testing as opposed to the second half. Further, experience levels did not appear to influence T600 performance and there were no significant interactions.

In terms of subject performance as contrasted with T600 performance the indication was that operator loading had a significant effect on subject verbal error rate and experience level was also a significant factor. A surprising aspect of experience was the observation that "little experience" level subjects had a higher error rate than "no experience" subjects. No explanation was offered for this observation and in fact these

findings may be spurious. This observation needs further study to examine the potential reasons for this finding.

In general Poock and Armstrong confirmed the following:

(1) Operator mental loading affected performance (recognition error rates were 23% greater with loading than under no mental loading condition).

(2) Performance appeared to be sensitive to trial half (i.e., recognition performance during the first 2.5 minutes differed from the second 2.5 minutes).

(3) T600 recognition errors were not influenced by experience level. This finding differs from the previous observations of subject performance where initial findings suggested an experience factor.

In summary, Poock and Armstrong's results were significant in that overall performance seems to be affected by mental loading. This finding could be extremely important in that the relationship between mental loading and performance could be indicative of what might be expected in operational military situations. Therefore, the area merits additional research to determine the extent of mental loading influence, possible relationships between nature of mental loading and performance, etc.

Equipment

Very few experimental efforts for voice recognition at NPS have dealt with equipment modification. Most studies have taken the existing system and investigated it's capability under various tasking or environmental conditions.

Schwalm (1982) recognized that under certain operational military environments certain equipment modifications may be required for satisfactory functioning. He postulated that under conditions where several operators were performing a task, each using a separate recognizer, the potential for confusion could be quite high and recognition error potential increased.

Schwalm suggested that one method for possibly decreasing errors in such a multioperator environment would be the addition of a mechanism whereby speaker commands could be directed to the microphone and further, provide a method for reducing the possibility of recognizible sounds or utterances to be released to the surrounding environment. He suggested the addition of a "mask" to currently available systems as one potential method for improving overall system performance in the multioperator environment.

The expressed objective of Schwalm's experiment was to examine the accuracy of an available voice recognition system with the addition of a "stenographer's" mask as compared to the conventional input device.

Initially 36 subjects (32 males and 4 females) participated in the study. However, as a result of the duration of the experiment and resultant scheduling problems, data was analyzed from 18 subjects (14 males and 4 females).

Equipment consisted of two T600 voice recognition systems. Both systems were capable of handling 256 discrete utterances. Three input methods were involved in Schwalm's study. First, a

conventional input device (SM10 boom microphone mounted on a headset). This is the normal input device for the T600. Second, a stenographer's mask with a microphone supplied by the manufacturer. The third input system consisted of a stenomask fitted with SM10 microphone.

Training of the speech recognizer was the standard process of 10 training trials per utterance. Testing consisted of two passes of the entire vocabulary on each of three successive days. Therefore six testing trials were run for each subject under each of the mask conditions.

In terms of total errors (misrecognition and nonrecognition) there was a significant mask effect. Results indicated a significant difference between no masks and both mask conditions. No difference existed between the mask conditions.

In terms of nonrecognition, no significant differences were observed. However, for misrecognition a significant difference was observed between no mask and both the original and the Shure mask.

One interesting observation was the fact that performance deteriorated over trials. This was true of total errors and misrecognitions. The author was unable to attribute these observations to any specific event and therefore considered these observations to be spurious. This assumption may or may not be valid and certainly warrants further consideration.

Schwalm also considered the potential influence of experience with masks and experience with microphones on

performance. Subjects were divided into two groups (high experience and low experience) for both mask and microphone use experience. Results suggested that mask experience was a significant variable in performance. Differences were observed between the no mask condition and both mask conditions in the group with low previous experience. In the high experience group, significant differences were observed between the no mask condition and the original mask condition.

In terms of microphone experience differences were observed between the no mask and the Shure mask condition for the low experience group and between no mask and the original mask for the high experience group.

Schwalm's effort is significant in that many military environments may involve the use of masks in the operational setting. That fact that a slight (3.5 percent) increase in errors between no mask and the average of the two masked conditions suggests a potential for performance degradation in performance. Admittedly the degradation observed was slight (performance with the mask was 94.7 percent correct recognition). The observation suggests the possibility of certain operational environments introducing perturbations that when considered in combination with other factors may degrade performance sufficiently to warrant new mask design configuration. Further, the observation that experience may be a factor in performance suggests the possibility of overcoming any degradation through training. Future efforts might consider the interaction between mask and experience/training in a simulated operational setting.

Summary

The research efforts on voice recognition are impressive in suggesting the feasibility and potential utility of voice as an input mechanism for man-machine systems. Obviously as for any technological advance additional research is suggested by the completed work. On the basis of current findings it would seem that one area in need of pursuit is the possible influence of task specificity. Similar task types occasionally produced somewhat contradictory results. Questions arise as a result of these observations as to whether the observed differences were the result of task specific conditions or were they the result of subtle experimental design questions?

It is also obvious that additional work on the potential influence of varied environmental factors be pursued. The environments explored (e.g. noise) need to be expanded upon and other potential physical environmental factors (e.g. vibration) need to be explored. In addition, it may be that various psychological environments may contribute to performance and these are areas of merit.

One area not considered and of potential importance is the general area of acceptance of the system by users and managers. Mercherikoff and Mackie (1970) suggested that operational military personnel frequently fail to totally accept and occasionally totally reject innovation in operational equipment and procedure. This resistance to change is not unique to the military and in fact appears to be a universal human characteristic. In the military, nonacceptance of new

equipment/technology can result in system failure or system rejection.

Based on subjective questioning of "users" in the artificial laboratory environment, rejection would not appear to be a significant problem. It must be realized that the subjects involved in experimentation are not really users in the operational sense and therefore the data collected may well be inapplicable. This area of technology acceptance should be considered in further research efforts.

In summary, the potential of using speech recognition in the military environment is impressive. Efforts conducted at the Naval Postgraduate School have been successful in suggesting the variety of tasks and environments in which speech recognition is an effective input device. Research in the operational environment would appear to be a most appropriate next phase of a total research program. Such efforts should consititute "research" as opposed to "demonstrations", however. Attempts should be made to examine the utility and effectiveness of voice in valid operational settings.

Further, the entire area of acceptance, fortunately with a system as novel as voice recognition, would appear to be of considerable merit. For even if the system is effective and offers definite advantages over more traditional systems, such advantages are lost if management and/or the individual user is unwilling to maximize the benefits and take advantage of the power of the system.

Obviously, considerable research needs to be accomplished before the potential and/or the limitations of voice recognition as a vehicle for human input to machine is realized. In fact, if one considers only those elements (i.e., equipment, environment, task and personnel) which have been suggested as determining the efficiency with which men interact with machine, it is obvious that some of the elements have received very little attention (e.g., environment). Individual "elements" need additional pursuit, as well as possible interaction between elements. Questions such as task specificity, different physical environments, training, etc. need to be addressed in future research. However, work already accomplished is certainly suggestive of the potential and can be considered as indicating that voice input is a reasonable and attractive alternative in many situations currently employing manual entry. Reliability of the system has been proven, now the question is one of specific application.

Bibliography

Armstrong, J.W. The Effects of Concurrent Motor Tasking on Performance of A Voice Recognition System. Masters Thesis, Naval Postgraduate School, Monterey, CA, 1980.

Armstrong, J.W. and Poock, G.K., Effect of Task Duration on Voice Recognition System Performance. NPS Technical Report No. NPS-55-81-017, Monterey, CA, 1981.

Batchellor, M.P. Investigation of Parameters Affecting Voice Recognition Systems in C^3 Systems. Masters Thesis, Monterey, CA, 1981.

Connolly, D.W. Voice Data Entry in Air Traffic Control. National Aviation Facilities Experimental Center, Report No. FAA-NA-79-20, August, 1979.

Drennen, T.G. Voice Interactive Systems: Applications and Payoffs, Voice Technology for Systems Applications Subgroup of the Department of Defense Human Factors Engineering Technical Advisory Group, Conference held 13-15 May 1980.

Elster, R. The Effects of Certain Background Noises on the Performance of a Voice Recognition System. NPS Technical Report. No. NPS-554-8-010, September 1980.

Jay, G.T. An Experiment in Voice Data Entry for Imagery Intelligence Reporting. Masters Thesis, NPS, Monterey, CA, 1981.

Lea, W.A. Value of Speech Recognition Systems In: Trends in Speech Recognition, W.A. Lea (Ed.), Englewood Cliffs, N.J.: Prentice Hall, 1980.

Lea, W.A. and Shoup, J.E., Review of ARPA SUR Project and Survey of Current Technology in Speech Understanding. ONR Speech Community Lab. Report, 1979.

Martin, T.B. and Welch, J.R., Practical Speech Recognizers and Some Performance Effectiveness Parameters In: Trends in Speech Recognition. W.A. Lea (Ed.), Englewood Cliffs, N.J.: Prentice-Hall, 1980.

McCormick, E.J. and Sanders, M.S., Human Factors in Engineering and Design. New York: McGraw Hill, 1982.

McSorley, W.J. Using Voice Recognition Equipment to Run the Environmental Simulator (WES), Masters Thesis, MPS, Monterey, CA, 1981.

Meister, D., Human Factors: Theory and Practice. New York: Wiley Interscience, 1971.

Mecherikoff, M. and Mackie, R.R. Attitudinal Factors in the Acceptance of Innovations in the Navy. Human Factors Research, Inc. Technical Report #784-1, Goleta, CA, 1970.

Neil, D. and Andreason, T., Examination of Voice Recognition System Ability to Function in a Bilingual Mode. NPS Technical Report No. NPS-55-81-003, February, 1981.

Pearson, R.G. and Byars, G.E., Jr., The Development and Validation of a Checklist for Measuring Subjective Fatigue. University School of Aviation Medicine, USAF, Report 56-115, December, 1956.

Poock, G.K. Experiments with Voice Input for Command and Control: Using Voice Input to Operate a Distributed Computer Network. NPS Technical Report No. NPS-55-80-016. April, 1980.

_____ A Longitudinal Study of Computer Voice Recognition Performance and Vocabulary Size. NPS Technical Report No. NPS-55-81-013. June, 1981.

Ruess, J.C. Investigation Into Air Launch Cruise Missile (ACLM) Flight Information Loading and Display Techniques During Flex Targeting Procedure. Masters Thesis, Monterey, CA, 1982.

Schwalm, N.D., Poock, G. K. and Roland, E. F. Use of Voice Recognition Equipment with Stenographer Masks. Naval Postgraduate School Technical Report NPS55-82-028, 1982.

Wolfe, C.D. and Taggart, J.C., Speech Recognition as an Input Medium for Preflight in the P3C Aircraft. Masters Thesis, NPS, Monterey, CA, 1981.

Distribution List

	NO. OF COPIES
Defense Technical Information Center Cameron Station Alexandria, VA 22314	2
Library, Code 0142 Naval Postgraduate School Monterey, CA 93943	2
Library, Code 55 Naval Postgraduate School Monterey, CA 93943	2
Professor D.E. Neil Code 55Ni Naval Postgraduate School Monterey, CA 93943	150
Dean of Research Code 012A Naval Postgraduate School Monterey, CA 93943	1